International Journal of Signal and Image Processing Issues
Vol. 2015, no. 1, pp. 27-36
ISSN: 2458-6498
Copyright © Infinity Sciences

A Study on Automatic Speech Recognition

Kumuthaveni Sasikumar[1], ChandraRavichandran[2]

[1] College of Arts & Science, Coimbatore-49, Tamilnadu, India
[2]Bharathiar University, Coimbatore-46, Tamilnadu, India
Email: *kumuthaveni@gmail.com*

Abstract -Automatic Speech Recognition (ASR) is an essential component for applications based on human-machine interfaces. Even after years of development, ASR remnants as one of the crucial research challenges like language variability, vocabulary size and noise. There is a need to develop Human-machine interface in native languages. In this regard, review of existing research on speech recognition is supportive for carrying out further work. Intend of Automatic speech recognition system requires cautious interest to the issues such as category of speech types, feature extraction, pattern classification etc. Here the paper presents a study on typology in ASR, the various phases involved, a brief description on each phase, basic techniques that make up automating speech recognition and different approaches to gain ASR. As an account of the brief study the paper shows enhanced precision results and good accuracy. The paper also displays a swot on speech recognition applications evoking research developments.

Keywords: *Automatic speech recognition; Acoustic model; Pattern classification; Hidden Markov Model; Linguistic model.*

I. INTRODUCTION

Human language is the method of translation of thought into physical output that enables humans to communicate [1]. The people are comfortable by speaking directly to machines than depending on other interfaces like keyboards and pointing devices. These interfaces require educated knowledge, patience and good hand-eye coordination for effective usage.

Speech recognition makes the task of converting any speech signal into its orthographic representation [3]. It is a technology that allows spoken input into systems. It means users talking to computers and computers recognizing it correctly. It needs analysis and conversion of the speech signal into the basic units of speech like phonemes or words. Then it interprets the elementary units for correction of words for linguistic processing [4]. The Speech recognition systems combine the interdisciplinary technologies from signal processing, pattern recognition, natural language and linguistics into a merged statistical framework.

II. TYPES IN AUTOMATIC SPEECH RECOGNITION

Automatic speech recognition is classified based on two techniques. They are based on the systems that understand speech and the speech utterances that the systems recognise.

A. Type of Systems in Automatic Speech Recognition

There are three types of systems in Speech recognition namely speaker dependent, speaker independent and speaker adaptive systems. A speaker dependent system is the system that is developed to operate for a single speaker. A speaker independent system is developed to operate for any speaker of a particular type and a speaker adaptive system is developed to adapt its operation to the characteristics of new speakers.

B. Types of Utterances of Speech in Automatic Speech recognition

Types of speech recognition are classified based on the types of utterances the system is capable to recognize. An utterance is the pronunciation (speaking) of a word or words that represent a single meaning to the computer. Utterances can be a single word, a few words, a sentence, or even multiple sentences[5]. These types are classified as the following:

1. Isolated Word Speech Recognition

The isolated word recognition systems characterise each word with pauses before and after it, so that end-pointing techniques can be used to identify word boundaries reliably [6]. This system is capable of a single word or single utterance at a time. These systems have "Listen/Not-Listen" states, where the speaker is required to wait in between utterances and in that gap processing is performed.

2. Connected Word Speech Recognition

Connected words are the words similar to the isolated words which allow separate utterances to be spoken together with a least pause between them.

3. Continuous Speech Recognition

Continuous speech recognition permits users to speak almost naturally where the system determines the content. Speech recognition with continuous speech capabilities are most difficult to create because they apply special methods to determine utterance boundaries.

4. Spontaneous Speech Recognition

Spontaneous Speech can be described as the speech that is natural sounding and has not been prepared or not rehearsed. An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as fillers, hesitations such as, "ums" and "ahs", pauses and stutters [7].

III. PHASES IN AUTOMATIC SPEECH RECOGNITION

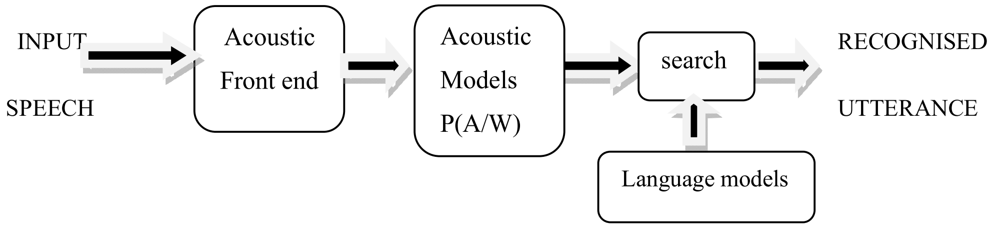

Figure-1: *Basic Model for Speech recognition*

Automatic speech recognition (ASR) systems follow a well-known statistical hypothesis as parameter conversion of speech signals at front-end (encoding) and recognition of the acoustic parameters (decoding) at the back end. Here the signal processing front end generates features from given acoustic signal and the back end recognizer uses these features to classify the input signals to output the recognized hypothesis [8].

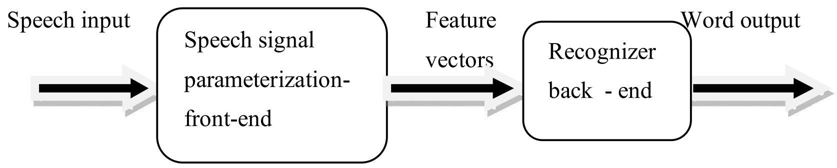

Figure-2: *Architecture of ASR*

A. Front-End Analysis for ASR

In Pre-processing of speech, the acoustic signal received as input to speech recognizer is converted into a sequence of acoustic feature vectors. It undergoes the steps such as pre-emphasis, frame blocking and windowing, feature extraction [9].

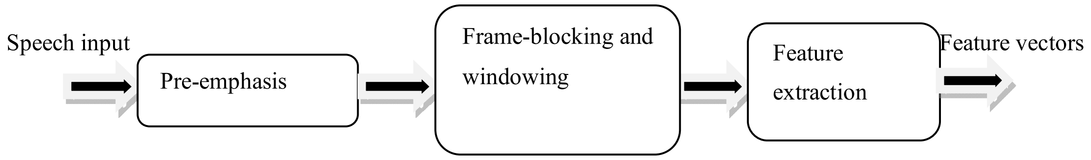

Figure-3: Front-end Analysis for ASR

1. Pre-Emphasis

The Pre-emphasis filter is used to flatten the speech spectrum before the spectral analysis to compensate the high-frequency part of speech signal that was suppressed during the human sound production mechanism. The most commonly used pre-emphasis filter is FIR (Finite Impulse Response). The pre-emphasis filter aids the spectral analysis process in modelling the perceptual aspects of speech spectrum [9].

2. Frame Blocking and Windowing

The speech signal is divided into a sequence of frames for independency and representation by a single frame or feature vector. Here each one of the frame is possible of possessing a stationary behaviour with frame blocking (20-25ms) window applied at 10ms intervals with frame rate of 100 frames/s[1].

3. Feature Extraction

Feature extraction techniques in speech recognition aims at separating source and filter components of the input signal and extraction of useful and relevant information from the speech frames providing arguments of these as feature parameters or vectors [3]. The filter bank analysis and cepstrum generation are important steps in Feature extraction.

The common parameters are MFCC (Mel Frequency Cepstral coefficients) and LPC (Linear Prediction coefficients)

a. MFCC

In MFCC the speech spectrum is passed through a Mel filter bank and the filtered output energies are log-compressed and transformed to cepstral domain by Discrete Cosine Transformations (DCT).

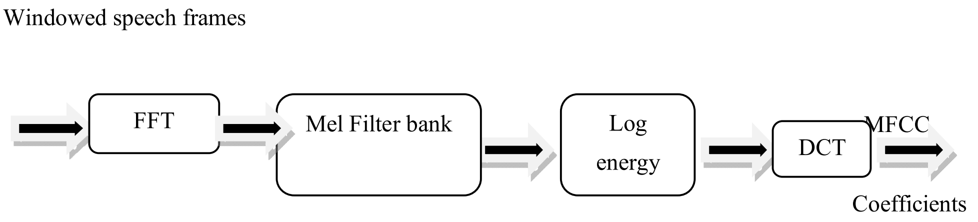

Figure-4: *MFCC feature vector*

b. LPC

In LPC, linear prediction filters attempt to predict future values of the input signal based on past signals. The frame of windowed signal is auto correlated and then transformed to the gain and autoregressive coefficients by using Levinson-Durbin recursion [9].

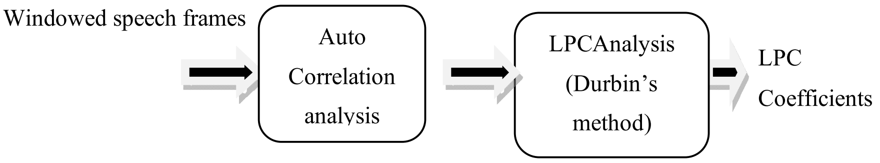

Figure-5: *LPC feature vector*

B. Back End Process in Speech Recognition

Back end of ASR uses a set of acoustic model and language model to decode the input string of words. The recognition process generally use pattern recognition approaches applying training and testing of the systems[3]. At the training period acoustic and language models are developed and posted as knowledge sources to decoding. The acoustic model plays a major role of mapping the sub word unit to acoustic observation. The language model introduces the linguistic restrictions in the language and permits recovery of illogical phoneme sequences. A phonetically rich database is required to train the acoustic models.

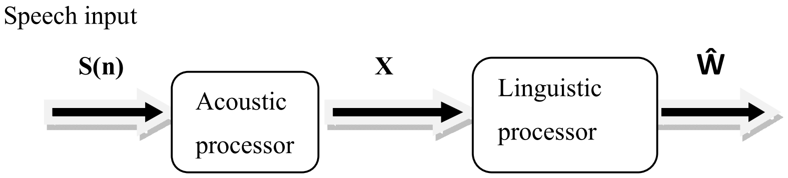

Figure-6: *Back-end model for Speech Recognition*

1. *Acoustic Processor*

Among the different acoustic models, HMM is the most commonly used technique due to its efficient algorithm for training and recognition. The vital power of the HMM is that it combines modelling of stationary stochastic processes and the temporal relationship among the processes together in a well defined probability space. The acoustic processor performs spectral analysis producing the spectral representation $X = (X_{1,} X_{2 \ldots} X_L)$ for L frames of speech signal [10].

2. *Linguistic Processor*

The Linguistic processing or Language modelling is a former knowledge description of a language. The goal of the Linguistic processor is to afford an estimate of the probability of a word sequence W for the given recognition task. This knowledge is free from an utterance to be recognised. The knowledge about a language can be expressed as words or word sequences that are possible and how often they occur. The Linguistic processor serves the purpose to resolve the ambiguities which the acoustic processor is not able to handle. The ASR systems use n-gram language models to conduct the search by predicating the possibility of the nth word using n-1 previous words. If we suppose that W is a specified sequence of words i.e., $W=w_{1,} w_{2 \ldots} w_n$

$$P\,(W) \;= P\,(w1, w2, \ldots\ldots.. wn\,) \tag{1}$$

$$= P\,(w1\,)\, P(w2|w1\,)\, P(w3|\, w1, w2\,) \ldots P(wn|\, w1\, w2\,, \ldots\ldots.. wn-1) \tag{2}$$

$$P(W) \;= \prod_{i=1}^{N} P\,(wi|\, w1\, w2\,, \ldots\ldots.. wi-1) \tag{3}$$

The familiar and reasonable n-gram models are tri-grams (n=3), where $P(w_3|\, w_1, w_2)$ is modelled for words w_1, $w_2$and w_3 and bi-grams(n=2), where $P(w_2|\, w_1)$ is modelled for words w_1 andw_2[8] .

The decoding process with trained acoustic and language models is often referred to as search problem. The linguistic processor deals with solution for maximum likelihood sentence, $\hat{W}$, best matching X using the Bayesian formulation:

$$\hat{W} = \arg\max_W P(W|X) \tag{4}$$

$$\hat{W} = \arg\max_W \frac{P(X|W)P(W)}{P(X)} \tag{5}$$

$$\hat{W} = \arg\max_W P_A\,(X|W)\, P_L(W) \tag{6}$$

The maximisation of equation (4) is converted to (5) using Baye's rule and the denominator P(X) is independent of the input sentence W, which can be removed in (6) leading to a three step solution. Here P_A (X|W) denotes the label of the acoustic model of P(X|W) and P_L(W) is the label for the language model for P(W) describing the various word combinations[10].

IV. APPROACHES TO SPEECH RECOGNITION

A. Acoustic Phonetic Approach

The Acoustic phonetic approach proposes that the spoken language consists of finite, distinctive phonetic units that are characterised by speech signal over time. It is assumed that though the acoustic properties of speech signal are highly variable, the rules governing the variability can be learned and applied in practical situations.

The first step in acoustic phonetic approach is the combination of spectral analysis with feature detection that converts the spectral dimensions to a set of broad acoustic properties that are developed in the speech signal over time. The second step is segmentation and labelling phase. Here it involves segmenting the speech signal into discrete regions where the acoustic properties of the signal represents one or several phonetic units and attaching one or more phonetic labels to each segmented region according to acoustic properties[5].

The next step tries to recognise a correct word or words from the phonetic labels determined from the first step that satisfies the constraints of speech recognition task(i.e., the words are drawn from given vocabulary, the word sequences makes a syntactic sense and has semantic meaning, etc.)

Thus the acoustic phonetic approach to speech recognition decodes the phoneme lattice (representing a sequential set of phonemes that are likely matches to spoken input speech obtained as a result of segmentation and labelling phase) into a word string that includes every instant of time in one phoneme lattice and the word obtained is valid according to rules of English syntax.

B. Pattern Recognition Approach

The pattern recognition approach unlike acoustic phonetic approach use speech patterns directly without explicit feature determination and segmentation. The method undergoes two steps.

The first step undergoes training of speech patterns i.e., speech knowledge is brought into the system and the machine learns the acoustic properties of the speech class and characterisation of speech through training. In the training, enough versions of pattern to be recognised are included in the training set that will help to adequately characterise speech patterns. This type of characterisation of speech through training is called pattern classification because the machine learns which acoustic properties of speech are reliable and repeatable across all training tokens of the pattern.

The second step of pattern recognition is the pattern comparison stage that does direct comparison of the unknown input speech with each possible pattern learned in the training phase. Here the measure of similarity between test pattern and reference pattern is computed using local distance measure (defined as the spectral distance between two well defined spectral vectors) and global time alignment procedure also called dynamic time warping algorithm which compares different rates of speaking of the two patterns [9].

Finally decision logic, where the reference pattern similarity scores decide which pattern best matches the unknown test pattern. The pattern recognition approach is used as a method of choice for speech recognition.

C. Artificial Intelligence Approach

Artificial intelligence (AI) approach is a hybrid of the acoustic phonetic approach and pattern recognition approach which combines the ideas and concepts of both methods. Here this approach attempts to mechanise the procedure of recognition by the way a human applies his intelligence in visualising, analysing and deciding the best acoustic features. In general AI approach to segmentation and labelling is to augment the acoustic knowledge with phonemic, lexical, syntactic, semantic and even pragmatic knowledge. There are three ways to integrate knowledge with speech recognition [9]. They are,

a) Bottom-up approach
b) Top-down approach
c) Blackboard approach

In Bottom-up approach the lowest level processes namely feature detection and phonetic decoding precede the higher level processes i.e., lexical decoding and language model in a sequential manner.

In Top-down approach the language model generates word hypothesis that are matched against the speech signal after wards the syntactically and semantically meaningful sentences are built upon the basis of the word match scores.

In Blackboard approach all knowledge sources are measured autonomous. A hypothesis and a test concept act as the essential means of contact among knowledge sources. Here every knowledge source is data driven based on the occurrence of patterns on the black board that matches the templates specified by knowledge sources.

D. Neural Networks and Their Approach to Speech Recognition

The neural network is an AI concept of automatic knowledge acquisition and adaptation. The form of the neural model is the feed-forward connectionist network. The neural network is a parallel distributed processing model that is a dense interconnection of simple, non-linear, computation elements. The acoustic speech signal is analysed by an "ear model" that stores spectral information about the speech input signal in the sensory information store. Both the long-term (static) memory and short-term (dynamic) memory are available to the feature detectors [9]. The final output of the system is obtained after several stages of refined feature detection and interpretation of the input information.

V. TECHNIQUES USED IN SPEECH RECOGNITION

A. Dynamic Time Warping (DTW)

The Reference-Pattern Model is a form of a store of reference patterns representing the voice-pattern space. To oppose misalignments, arising from change in speaking rate, for example, temporal alignment using dynamic time warping (DTW) is often applied during pattern matching [10].

DTW is used to find the matches between two given sequences (e.g. time series) may be audio, video or graphics indeed that are capable of turning into linear representation. ASR is a well known application to manage with different speaking speeds. Here the similarity between sequences independent of non-linear variations in time dimension is measured by warping the sequences non-linearly in time dimension [9]. This sequence arrangement method is frequently used in the framework of hidden Markov models. An example of the limitations forced on the matching of the sequences is on the monotonicity of the mapping in the time dimension. On comparing with other pattern matching algorithms continuity is less important in DTW than in other algorithms; above all DTW is an algorithm principally suited to matching sequences with missing information, given there are lengthy enough segments for matching to occur. The optimization process is performed using dynamic programming, hence the name.

B. Vector Quantization (Vq)

The reference patterns in pattern matching may represent a compressed pattern space, on average obtained through vector averaging. Compressed-pattern-space approaches seek to reduce the storage and computational costs associated with an uncompressed space. They include VQ models [11]. A conventional VQ model consists of a collection (codebook) of feature-vector centroids. Vector quantization is a method of efficiently representing the time varying spectral characteristics of the speech signal. On comparing the information rate of vector representation to that of raw speech signal, spectral analysis representation has reduced the information rate. The raw signal of information rate 160,000bps is used for storage in an uncompressed format. On spectral analysis of the same signal, with dimension p=10 using 100 spectral vectors per second representing each spectral component by a 16-bit precision, 16,000bps is used for storage resulting in a reduction of 10-1 over raw signal. VQ is the concept of building a code book of analysis vectors with more code words more than the set of phoneme [9]. Then to represent an arbitrary spectral vector on VQ there is a need for a 10-bit number for the index of the code book. On an assumption of a rate of 100 spectral vectors per second, a total bit rate of about (10*100) 1000bps is required to represent spectral vectors of speech signal which is about 1/16th the rate required by continuous spectral information in speech signal.

C. Hidden Markov Models (HMM)

HMMs are generative data models that suits good for the statistical modelling and identification of sequential data, such as speech. HMM implants two stochastic mechanisms. One mechanism is a Markov chain of hidden states that models the sequential evolution of observations. The other mechanism is a set of probability distributions of observations where each state has one distribution. A discrete or a continuous function represents this distribution. This mechanism divides HMMs into discrete-density HMMs (DHMMs) and continuous-density HMMs (CHMMs), respectively [10]. HMMs are learnable finite stochastic automates. An HMM is a statistical model of a class of speech signals that is capable of assessing the probability that a given input sequence belongs to the class on which it was trained. The class can be a word, phoneme or sub-word models which can be used as building blocks to produce word or multi-word recognisers. Here the difficulty of finding an interpretation amid the many possible interpretations becomes supreme. It is at this point that we commence to include models of the language we are recognising to help to limit the search and also we consider sequence of phonemes or word only those which fit our models of the language and the task.

The doubly stochastic processes namely states and transition probabilities together make up the Hidden Markov Model. The model is called "hidden" because the states of the chain are outwardly not visible [5]. The transition probability produces emission visible at each instant, depending on a state-dependent probability distribution.

VI. USES AND APPLICATIONS

ASR technology has been applied in diverse fields commencing from telephonic environment followed by educational sector, entertainment, domestic, military, artificial intelligence, medical, agriculture, general, dictation, translation etc. and can list out various applications available under ASR.

Table 1. *Uses and Application of ASR*

Domain	**Application**
Telephony	Directory enquiry without operator assistance.
Education	Teaching foreign languages, educate physically handicapped students
Domestic	Commanding the electronic devices instead of using buttons.
Military	Training air traffic controllers, fighter aircrafts, helicopters, battle management.
Artificial intelligence	Robotics
Medical	Health care, Medical Transcriptions.
Agriculture	To intimate the agriculturalists, the knowledge about the varying market rates of the farm products, seasonal crops, proper pesticides etc.,
General	Multimedia interacting, court reporting, grocery shops
Dictation	Replacing menu systems by commanding computers.
Translation	Advanced applications that translates from one language to another.

VII. VII.DISCUSSION

ASR systems face challenging opportunities under many slots. The microphones design needs gradation to adapt rapidly to the changing background noise levels, different environments and discarding of extra noise during recognition. ASR systems performances degrade based on noise and reverberation environment during speech recording of microphones for recognition. The ASR needs improved performance to allow batch processing of speech recognition. Performance of speech recognition is a great challenge to make the entire human population of

India to enjoy the benefits of internet evolution. The developing country like India where there are about 1670 dialects of spoken form, speech recognition technology has wider scope and application [2].

VIII. PERFORMANCE METRICS OF SPEECH RECOGNITION SYSTEM

Speech recognition system performance is generally measured by means of accuracy and speed. The word error rate (WER) measures accuracy and real time factor (RTF) measures the speed of speech recognition systems [5]. The single word error rate (SWER) and command success rate (CSR) are also used as accuracy measures.

The general problem of measuring performance is based on the fact that the recognized word sequence can have a different length from the reference word sequence (say the correct one). The WER is derived from the Levenshtein distance which works at the word level instead of the phoneme level [12]. This difficulty is solved with initial alignment of the recognized word sequence with the reference (uttered) word sequence using dynamic string alignment. Further the WER is determined from the equation below,

$$WER = (S + D + I)/N \tag{7}$$

Here S, D, I, N where the number of substitutions, deletions, insertions, N the number of words in the reference. Sometimes Word Recognition Rate is used to measure the performance of speech recognition and is defined in the equation below,

$$WRR = 1 - WER = [N - (S + D + I)]/N = (H - I)/N \tag{8}$$

Where WRR is the word recognition rate and H is N-(S+D), the number of correctly recognised words.

$$RTF = Decoding\ time/Speech\ duration. \tag{9}$$

IX. CONCLUSION

Speech recognition is an amazing area for research studies. The speech recognition is being used in a variety of assistive contexts, including home computer systems, mobile telephones, and various public and private telephony services. The system has found many applications in various fields like health care, military, business, legal transcription, etc. Speech recognition is a technology that has turned out to be really popular over the years and presently has a wide range of applications. There are still contents to be unwrapped under speech recognition technology. The different approaches to speech recognition under different techniques makes it research area broadening.

REFERENCES

[1] Benesty.J and Sondhi.M.M. and Huang.Y. *Natural Language Understanding*. Springer Hand Book of Speech Processing. 2008, p. 617-626.

[2] CINI Kurian. A Survey on Speech Recognition in Indian Languages, *International Journal of Computer Science and Information Technologies. 2014, vol.* 5, no p. 6169-6175

[3] CHANDRA, E. et AKILA, A. An Overview of Speech Recognition and speech synthesis algorithms, *Int.J.Computer Technology & Applications. 2012, vol.3, no* 4, p.1426-1430.

[4] JUANG, B.H. The Past, Present and Future of Speech Processing, *IEEE Signal Processing Magazine*, 1998, p. 24-48.

[5] ANUSUYA, M. A. et KATTI, S.K. Speech recognition by machine – a review, *International Journal of Computer Science and Information Security*, 2009, vol 6, no 3, p. 181-205.

[6] RABINER, Lawerence R. et LEVINSON, Stephen E. Isolated and connected word recognition--Theory and selected applications. *Communications, IEEE Transactions on*, 1981, vol. 29, no 5, p. 621-659.

[7] SATHYA, Ms MA Josephine. A STUDY ON TRAINING, COMPARISON OF WORDS IN AUTOMATIC SPEECH RECOGNITION USED FOR FLUENCY DISORDER THERAPY–STUTTERING. 2014.

[8] AGGARWAL, R. K. et DAVE, M. Integration of multiple acoustic and language models for improved Hindi speech recognition system. *International Journal of Speech Technology*, 2012, vol. 15, no 2, p. 165-180.

[9] RABINER, Lawrence et JUANG, Biing-Hwang. Fundamentals of speech recognition. 1993.

[10] RABINER, Lawrence et JUANG, Biing-Hwang. Historical Perspective of the Field of ASR/NLU. In : *Springer Handbook of Speech Processing*. Springer Berlin Heidelberg, 2008. p. 521-538.

[11] CHIBELUSHI, Claude C., DERAVI, Farzin, et MASON, John SD. A review of speech-based bimodal recognition. *Multimedia, IEEE Transactions on*, 2002, vol. 4, no 1, p. 23-37.

[12] KARPAGAVALLI, S., DEEPIKA, R., KOKILA, P., *et al.* Automatic Speech Recognition: Architecture, Methodologies and Challenges-A Review. *International Journal of Advanced Research in Computer Science*, 2011, vol. 2, no 6.